ITEM RESPONSE THEORY AS A TOOL TO IDENTIFY STUDENT TOPIC ABILITIES

by

Braden Ohlsen

A thesis submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Master of Science

Department of Chemistry

The University of Utah

December 2017

The University of Utah Graduate School

STATEMENT OF THESIS APPROVAL

The thesis of **Braden Ohlsen**

has been approved by the following supervisory committee members:

Charles Atwood	, Chair	**8/16/17** Date Approved
Thomas Richmond	, Member	**8/16/17** Date Approved
Michael Morse	, Member	**8/16/17** Date Approved

and by **Cynthia Burrows**, Chair/Dean of

the Department/College/School of **Chemistry**

and by David B. Kieda, Dean of The Graduate School.

ABSTRACT

The goal of the Atwood group is to improve student success in general chemistry at the University of Utah. To accomplish this goal, we created a system of pretests that allows students to practice before the actual exam and analyze their pretest scores to assess their ability prior to taking an actual exam.

We developed a method to analyze an individual student's proficiency on the topics that make up a test utilizing Item Response Theory (IRT). This had not been done previously at the individual student level. We used this information to provide students with feedback on where to focus their studies. We hoped that the combination of extra practice, the opportunity for students to check their own progress and detailed feedback would result in improved outcomes on the exams.

After implementing this pretest system, equated student abilities on the midterm exams increased significantly compared to the previous year when the pretest system was not in place. The particular effect of the topic feedback was also studied by comparison with a control class but results were inconclusive.

CONTENTS

LIST OF FIGURES

Figures

INTRODUCTION

For students to improve their performance in chemistry, they first need to be aware of their own level of understanding (Dunlosky and Metcalfe, 2009). Not only do students occasionally need the proverbial wakeup call provided by a poor test score to put forth their best effort, but it has even been suggested that the metacognitive skills needed to assess one's own proficiency are directly linked to the actual skills of being proficient on a topic (Kruger and Dunning, 1999). Poor students are notoriously bad at assessing their own ability, leading to overconfidence going into tests and disappointment upon receiving their score.

Studies have shown that students overestimate their own ability when it comes to chemistry. This is particularly true for low-scoring students: the worse a student performs on a test, the greater the disparity between their actual ability and their self-perceived ability (Bell and Volckmann, 2011). When students were asked to predict how well they will do on an exam, high scoring students did so with a high degree of accuracy while those who scored poorly overestimated their own ability by a significant margin.

There are numerous studies identifying overall chemical topics that give students problems (Childs and Sheehan, 2009; Johnstone, 2006; Schurmeier, Atwood, Shepler, and Lautenschlager, 2010). Additional studies have identified what exactly makes topics such as bonding (Özmen, 2004), equilibrium (Chiu, Chou, and Liu, 2002), acidity (McClary and Talanquer, 2011) or properties of solution (Pınarbasi and Canpolat, 2003)

difficult. Item Response Theory (IRT) has even been used to identify areas that are especially difficult for students at a single school (Schurmeier, Shepler, Lautenschlager, and Atwood, 2011).

All of these studies, however, looked at the class as a whole, determining difficulties for the average student, not on an individual basis. While this approach is very useful at giving instructors information where they should focus their teaching, it fails to take into account the wide variety of errors and misconceptions found among individual students. One student might have a solid understanding of chemical nomenclature and quantum numbers but have deeply flawed conceptions regarding bonding. Another student might understand quantum numbers and bonding but have no idea how to name a simple compound. If both hypothetical students took a test that addressed these three concepts, they might end up scoring very similarly. Clearly though, one would not want to treat these students the same if one were teaching a class and the two students approached the instructor for help studying. He or she would sit down with each student as an individual to address the specific areas that they were struggling with and strive to improve their understanding of these areas. For a class of several hundred students, an instructor doesn't have enough time to meet and go through this process with every student individually. For a class of this size, one needs a way to analyze students' difficulties and inform them automatically. For this reason, we sought to design a method capable of determining student strengths and weaknesses at the individual level.

Item Response Theory

IRT is a paradigm of psychometric test analysis used for ability assessment (Ayala, 2009). It is an improvement compared to other analyses such as Classical Test Theory because it treats questions within a test as having different relative difficulties and discrimination abilities. This allows for more accurate scoring as well as more effective test administration. Different forms of the same test, such as used for the Graduate Record Examinations (GRE) or Scholastic Assessment Test (SAT), can be directly compared to one another (An and Yung, 2014). Another advantage of IRT is that it compares student ability and question difficulty on the same continuum, allowing one to predict how each individual student might do on a given question. We employed this characteristic of IRT for our analysis.

According to IRT, the probability of a student correctly answering a particular question is a mathematical function of student ability and several question parameters. IRT utilizes a logistic equation that yields an S-shaped curve, also called an Item Characteristic Curve (ICC) when probability is graphed against student ability. Various ICCs are shown in Figures 1, 2, and 3.

One-Parameter Model

The comparison between question difficulty and student ability provides the simplest form of this logistic model and is expressed mathematically as $P(\Theta,b) = \frac{e^{(\theta-b)}}{1+e^{(\theta-b)}}$ where $P(\Theta,b)$ is the probability of a correct response, Θ is student ability and b is question difficulty. The probability of a student correctly answering a question is dependent on the difference between their ability and the question difficulty. When

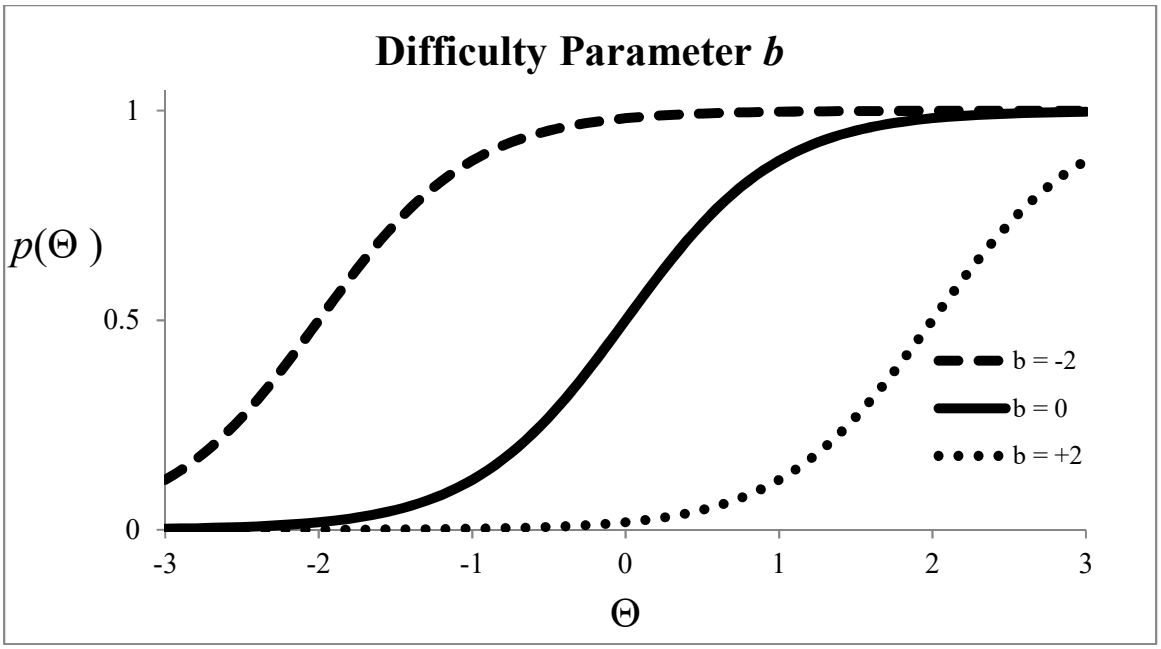


Figure 1: Item characteristic curves showing the effect of the difficulty parameter *b*. Easier questions have lower difficulties and therefore a higher probability of a student answering correctly.

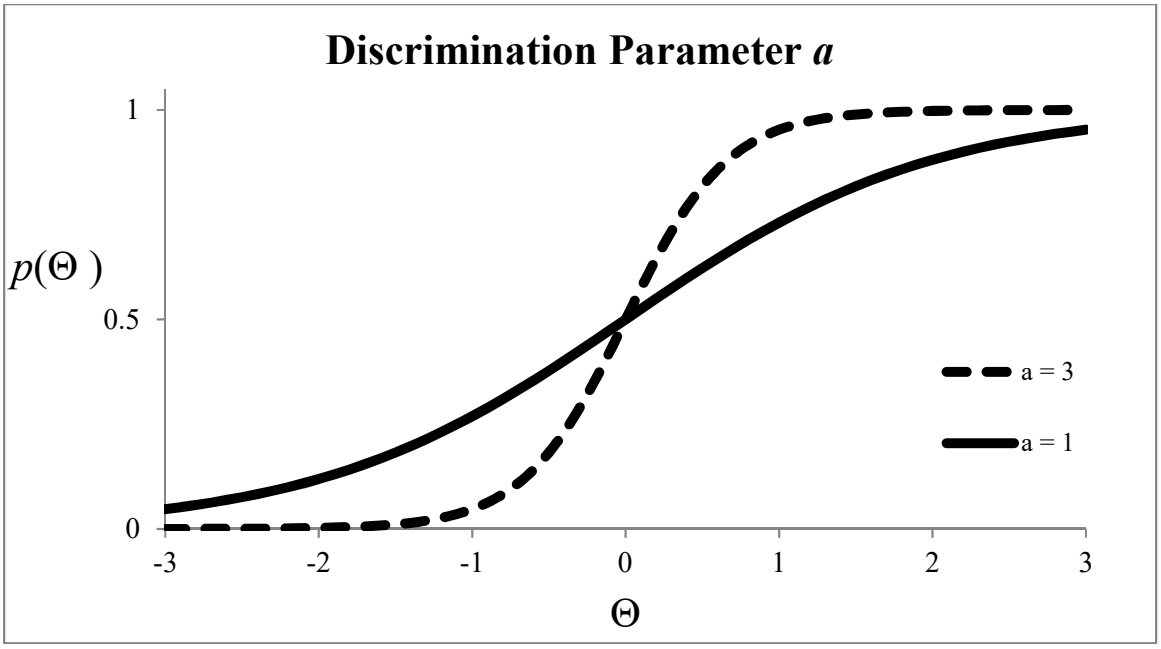


Figure 2: Item characteristic curves showing the effect of the discrimination parameter *a*. More discriminating questions have steeper slopes and can better discern a student's ability.

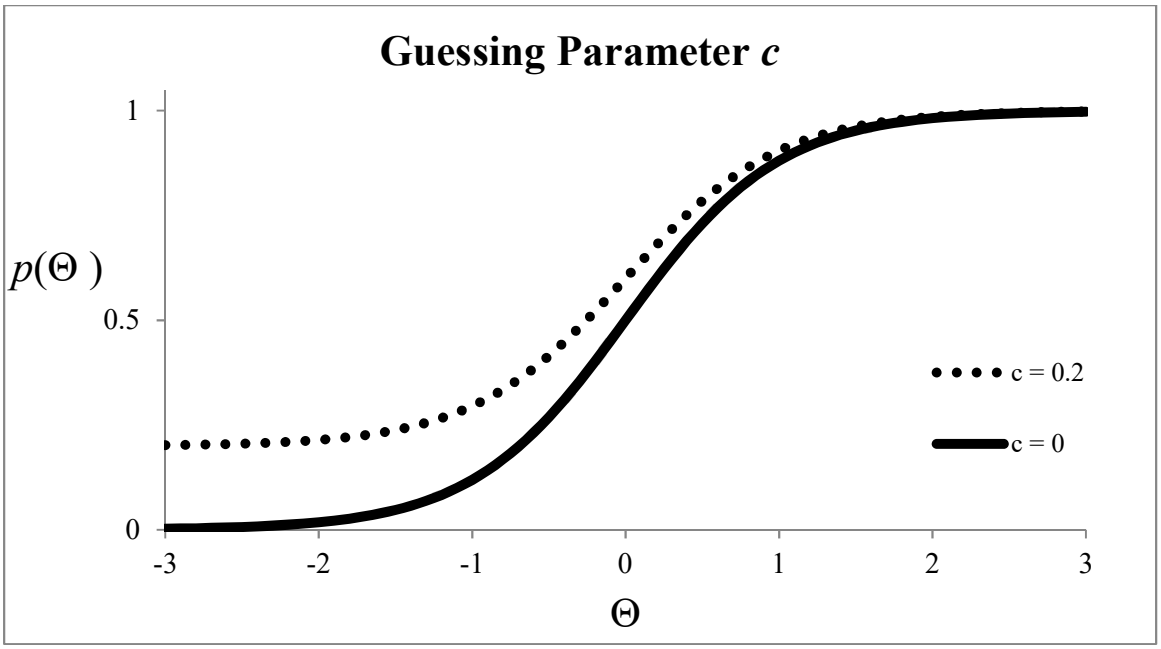


Figure 3: Item characteristic curves showing the effect of the guessing parameter *c*. The guessing parameter accounts for questions that a student has a chance of getting correct by chance alone.

student ability is higher or lower than question difficulty by a wide margin, the probability plateaus and approaches 1 or 0, respectively, as shown in Figure 1. This model, in which the probability function is based only on student ability and question difficulty, is referred to as the One-Parameter (or Rasch) Model. It is useful as a starting point for IRT analysis but limited compared to models that include additional parameters.

Two-Parameter Model

The Two-Parameter Model can potentially improve the data fit of the model by acknowledging that questions are unequal in their ability to differentiate between students of varying abilities. The second parameter in this model is thus a discrimination parameter, signified as *a*, and alters the logistic function to become $P(\Theta,a,b) = \frac{e^{a(\theta-b)}}{1+e^{a(\theta-b)}}$. Graphically, the *a* value determines the steepness of the S-curve with larger *a* values corresponding to steeper curves and more discriminating questions (Figure 2). From the vantage point of attempting to assess student ability, the discrimination parameter is a measure of how good a question is since high *a* values are better at differentiating between students of a similar ability level.

Three-Parameter Model

The third and final parameter that is involved in the Three-Parameter Model is the guessing parameter, designated *c* and shown in Figure 3. This parameter improves the fit relative to the One- or Two-Parameter models by providing a non-zero asymptote to the probability function. On a multiple choice question or even those requiring short answers, the probability of a truly hopeless student answering correctly is not equal to 0;

there's always the possibility they may guess the correct answer by chance. The guessing parameter takes this into account mathematically by altering the probability function to be $P(\Theta,a,b,c) = c + (1 - c)\frac{e^{a(\theta-b)}}{1+e^{a(\theta-b)}}$. This gives the Item Characteristic Curve a lower boundary and raises the inflection point of the curve.

METHODS

Prior to the Fall 2015 semester, a test review scheme was created with the goal of improving student performance on the midterm exams. It was then implemented in the first semester of general chemistry, CHEM 1210, in Fall 2015 and in the second semester, CHEM 1220, in Spring 2016.

Our primary goal was to determine a student topic ability (STA) for each student on each topic and provide them with this information, allowing them to most efficiently direct their studying to where they had the most to gain. With this in mind, we created a system of pretests during the week before each midterm exam. After each pretest, we analyzed it to determine the major topics from that pretest and then the STA for each student on each of those topics. This information was then converted into an easy format for the students to understand and distributed to them individually.

Pretests were created on the same testing platform that we had been using for midterm exams for several years, *Madra Learning* (Madra Learning, 2017). This had the advantage of allowing students to familiarize themselves with the exam format as well as providing us with detailed results for each student. For this purpose, pretests were designed to be similar in style and material to the exams. The pretests each contained 20 questions as compared to 25 for the exams and given a time limit proportional to that of the respective exam. Pretests were qualitatively selected to cover the same material as the exams but did not include identical questions.

To test the effect of the feedback itself, during the first semester of its implementation (Fall 2015), we utilized a control group within the pretest system. Since three different professors were each teaching the same course at the same time using the same syllabus, for simplicity's sake, we designated one of the three professors as the control. The roughly 200 students in that class (out of 900 overall) would take the same pretests as everyone else, but they would not receive the feedback detailing their relative strengths and weaknesses.

Hypothesis

We predicted that our pretest system would provide students with a way to practice prior to their exams and encourage low-ability students to study more, improving student ability on the exams compared to the previous year when such feedback was not provided. Furthermore, we predicted that by determining student topic abilities and providing students with that information, we would see an increase in the abilities of those students receiving the feedback relative to those who did not.

Topic Determination

To determine student topic abilities on a particular pretest, we first needed to determine the topics covered on that pretest. Initially, topics were determined qualitatively. This involved looking at the text of the questions themselves and sorting them into common groups. For example, the second test of General Chemistry 1 (CHEM 1210) covered nomenclature, bonding and some of the basics of quantum mechanics. This process was aided by the breakdown of topics in the class itself as part of the

textbook, lectures and homework. This was also intended to help students be familiar with the topic terminology that we included in our feedback.

Unidimensionality

For the IRT model to fit the student response data, it is assumed that the pretest questions within each category exhibit unidimensionality. Unidimensionality means that there is only one underlying trait (their ability on that specific topic) responsible for determining their responses. We can only determine STA's if the response data are unidimensional. The *NOHARM version 4* (Normal Ogive Harmonic Analysis Robust Method) software program (Fraser and McDonald, 1988) was used to assess unidimensionality.

NOHARM generates a residual matrix comparing the observed question covariances to the covariances generated by the one-dimensional model. Small residuals indicate good model fit. To quantitatively assess this, *NOHARM* generates two measures summarizing the residual matrix. First, *NOHARM* calculates the matrix's root mean square (RMS). If the RMS is less than $\frac{4}{\sqrt{N}}$ (where N is the number of students taking the pretest), it indicates a good model fit. Additionally, *NOHARM* generates Tanaka's goodness of fit index (GFI) (Tanaka, 1993). A GFI value of 1.0 would indicate a perfect model fit while values greater than 0.95 are generally considered to constitute a good fit.

IRT Analysis

IRT was used to determine each student's overall ability as well as their ability on each topic on the exam. The IRT analysis was conducted using *Bilog-MG3* software

(Bilog-MG 3.0, 2003) that employed Marginal Maximum Likelihood Estimation (MMLE). MMLE is an iterative process in which the question parameters are successively estimated, compared to the data being modeled and changed to correspond more closely with it. All of the student abilities are integrated to form a normal distribution by "chunking" similar-scoring students into quadratures. Each individual student's response to a question is assumed to be randomly sampled from its quadrature allowing question parameters to be estimated independently of student ability (Lord 1986).

Bilog-MG3 then estimates student abilities by fitting each student's responses to the calculated question difficulties using Maximum Likelihood Estimation (MLE). MLE calculates the likelihood (L) of a student with a particular ability producing their actual set of correct and incorrect responses. L is obtained by multiplying the individual question probabilities for a student with this ability: $P(\Theta, a, b, c)$ for a correct response and $1-P(\Theta, a, b, c)$ for incorrect. L is thus a function of student ability since each question's probability is also a function of ability. The maximum of this likelihood function is the student's ability and is calculated by iterating through a series of Θ values and calculating L for each (Ayala, 2009).

The three-parameter IRT model provided the best fit to our response data. Since the majority of our questions were multiple-choice, even the poorest students are expected to have some probability of answering a question correctly by chance. Therefore, we included the lower asymptote (sometimes called the "guessing parameter") in the model.

Student Topic Ability

STA's were calculated by treating each previously determined topic within a pretest as its own test. The student responses to only those questions that composed a topic, anywhere from 5 to 15 questions depending on the test and topic, were analyzed to determine ability for that topic

For example, a pretest towards the beginning of General Chemistry 1 might cover nomenclature, bonding and some of the basics of quantum mechanics. The questions involving nomenclature were separated from the rest and the matrix of results from those questions was analyzed. A typical matrix is shown in Table 1. First, this matrix was input into *NOHARM* and the questions in the topic were assessed for unidimensionality. Once unidimensionality was established, the matrix was input into *Bilog-MG3*, which calculated question and student parameters by MMLE and MLE, respectively, as described above. The calculated abilities thus reflected only student performance on nomenclature and represented the STA for nomenclature. This process was then repeated across all previously determined topics.

Feedback

After each pretest had been analyzed, topics determined and STA's calculated, the information was transmitted back to the students to guide their studying. To avoid the confusion of simply providing each student with their STA's, we converted these into a Likert scale making them easier for students to understand. Abilities greater than 1.5 (1.5 standard deviations above average) were considered as 'well above average,' 1.5 to 0.5 as 'above average,' 0.5 to -0.5 as 'average,' -0.5 to -1.5 as 'below average' and less than

Table 1: Result matrix for a single topic of a pretest. This topic contained nine questions (shown in rows) and the results from fifteen students are shown (columns). Ones indicate a correct answer while zeros indicate incorrect.

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9
I	1	0	1	0	0	0	1	1	1
II	1	1	1	1	1	0	1	1	1
III	0	0	0	0	1	0	0	1	0
IV	1	0	0	0	0	0	1	0	0
V	1	1	1	1	1	0	1	1	1
VI	1	1	1	1	1	0	1	1	1
VII	1	1	0	0	1	0	1	1	0
VIII	1	1	0	0	1	1	1	1	0
IX	1	1	0	0	1	1	1	1	0
X	1	1	0	0	0	0	1	1	0
XI	1	1	0	0	0	0	1	1	0
XII	0	1	0	0	0	0	0	1	0
XIII	1	1	0	0	1	1	1	1	0
XIV	1	0	0	0	0	0	1	1	0
XV	1	0	0	0	1	0	1	1	0

-1.5 as ‘well below average.’ After each pretest, the students were informed via an automatic email system of the Likert score of their STA’s.

IRT Equating

On each exam, there were 20 or 25 questions. Of these, 5 to 10 were conserved from one year to the next while the remaining questions differed. Because the majority of questions differed from year to year, to determine the change in student ability, it was necessary to account for these differences. This was done by IRT equating the two exams, which allowed the results on differing exams to be compared from one year to another and for a prediction to be made of how students would have scored if they were given the previous year’s exam.

This equating process was carried out using *IRTEQ* software. The 5 to 10 conserved questions acted as anchor points; by comparing the calculated question parameters for each year on these particular questions, it is possible to convert the question parameters and student abilities from one year to the IRT scale of the other year. The item characteristic curves (ICC) for the conserved questions were summed, resulting in a total characteristic curve (Figure 4). The total characteristic curves for each year were then compared and a linear regression was created to transform one total characteristic curve onto the other. This linear regression calculates two coefficients, in the form of $y = mx + b$, that can be used to convert student abilities from the scale of one year’s exam to that of another.

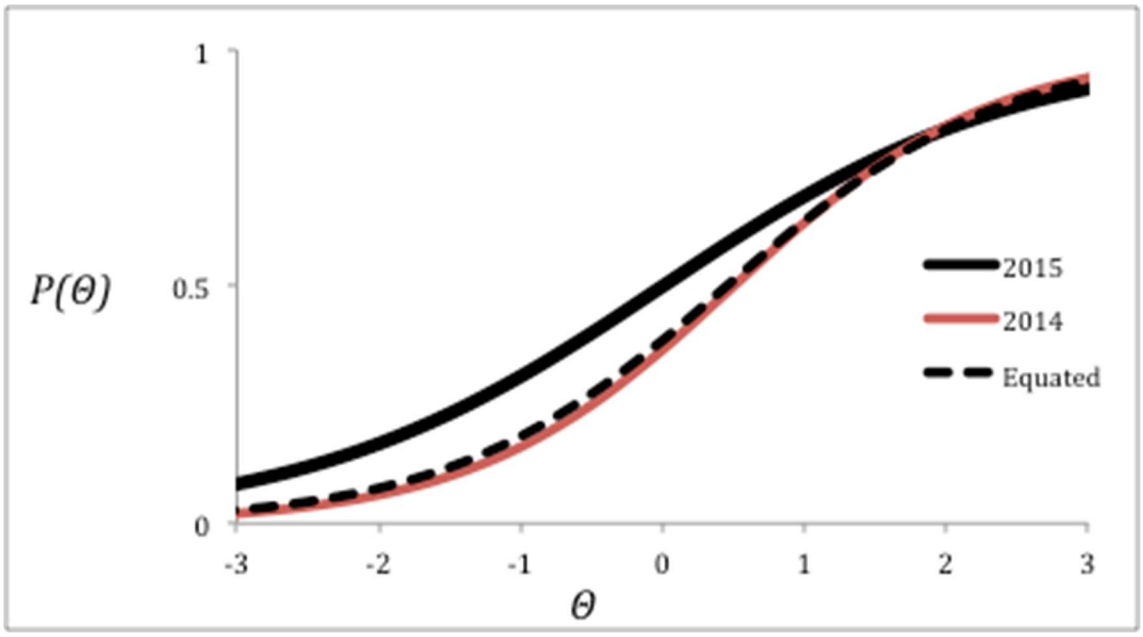


Figure 4: IRT total characteristic curves comparison for 2014 and 2015. The graph demonstrates the process of equating these curves for IRT ability comparison from 2014 to 2015.

RESULTS AND DISCUSSION

Unidimensionality

Before student topic abilities could be determined by IRT analysis, we needed evidence that the topics were unidimensional. An example of a residual matrix generated by *NOHARM* for one topic of a pretest is shown in Table 2. None of the individual residuals are particularly large and the RMS value of 0.0072 is well below the $\frac{4}{\sqrt{N}}$ value of 0.1489. This, combined with a GFI value of 0.9931, indicates a good fit for the unidimensional model for this pretest topic. Several of our pretest topics had GFI values as low as 0.95, but as this is still an acceptable value, and because the RMS values for these topics were still below $\frac{4}{\sqrt{N}}$, this led us to believe that all of our tested topics are unidimensional.

Student Topic Abilities

Once questions had been sorted into topics and after these topics were confirmed to be unidimensional, we determined the student topic abilities using IRT. An example of the STA's for twelve students on one pretest is shown in Table 3. As one might expect, a wide variety of student abilities were found. Some, like student I, were strong in all areas, while others, like student XII, had difficulties on all topics. Of particular interest were students like VI with a high ability in one topic (in this case naming) and a low ability in another (quantum). Providing these students with feedback could prove

Table 2: *NOHARM* –generated residual matrix. Small residuals indicate good data-model fit. RMS values less than $\frac{4}{\sqrt{N}}$ and GFI values greater than 0.95 are considered acceptable.

		Question Number					
		1	2	3	4	5	6
	2	-0.002					
	3	-0.012	-0.005				
Question	4	0.011	-0.008	-0.002			
Number	5	-0.011	0.005	0.012	0.006		
	6	0.01	0.006	0.002	-0.011	-0.006	
	7	0.006	0.001	-0.006	-0.003	-0.002	0.004

RMS = 0.0072
GFI = 0.9931

Table 3: A set of twelve student's topic abilities from a pretest. Abilities greater than 0 are above average while those less than 0 are below average.

Student	**overall**	bonding	naming	quantum
I	1.90	1.96	1.92	1.77
II	1.32	1.95	0.50	2.03
III	0.83	0.77	0.27	1.79
IV	0.78	0.98	1.90	-0.44
V	0.50	-0.80	1.25	1.48
VI	-0.16	0.13	1.23	-2.10
VII	-0.42	1.96	-1.83	-0.53
VIII	-0.95	-0.62	-0.72	-2.06
IX	-0.97	-1.65	-1.63	-0.41
X	-1.29	-1.98	-1.10	-1.54
XI	-1.64	-1.48	-1.99	-1.95
XII	-1.97	-2.03	-2.24	-2.20

particularly important because they have the most room for improvement if they focus their efforts on a single weak area.

Feedback

Once STA's had been determined after each pretest, feedback emails were sent to the students. Abilities greater than 1.5 were considered 'well above average,' 1.5 to 0.5 as 'above average,' 0.5 to -0.5 as 'average,' -0.5 to -1.5 as 'below average' and less than -1.5 as 'well below average.' The feedback each student in Table 3 received is shown in Table 4. For example, the aforementioned student VI would receive an email letting them know that they had scored average overall, average on bonding, above average on naming and well below average on quantum. This process was then repeated for each pretest during the week prior to each midterm exam.

Equated Abilities

Our hypothesis predicted that providing students with feedback on their pretest topic abilities would lead to better study habits and therefore higher exam scores. We used the exam score data from the school year of 2014 – 2015 as our control. We were looking for a difference in exam scores between that year and the following years (2015 – 2016 & 2016 – 2017) after we had introduced the pretest system. Although the courses in question, CHEM 1210 and 1220, were each taught several times during the year, we only studied 1210 in the fall semesters and 1220 in the spring because this sequence contains the vast majority of our students. The exams had the same time limit and contained the same number of questions from year to year, but the majority of questions

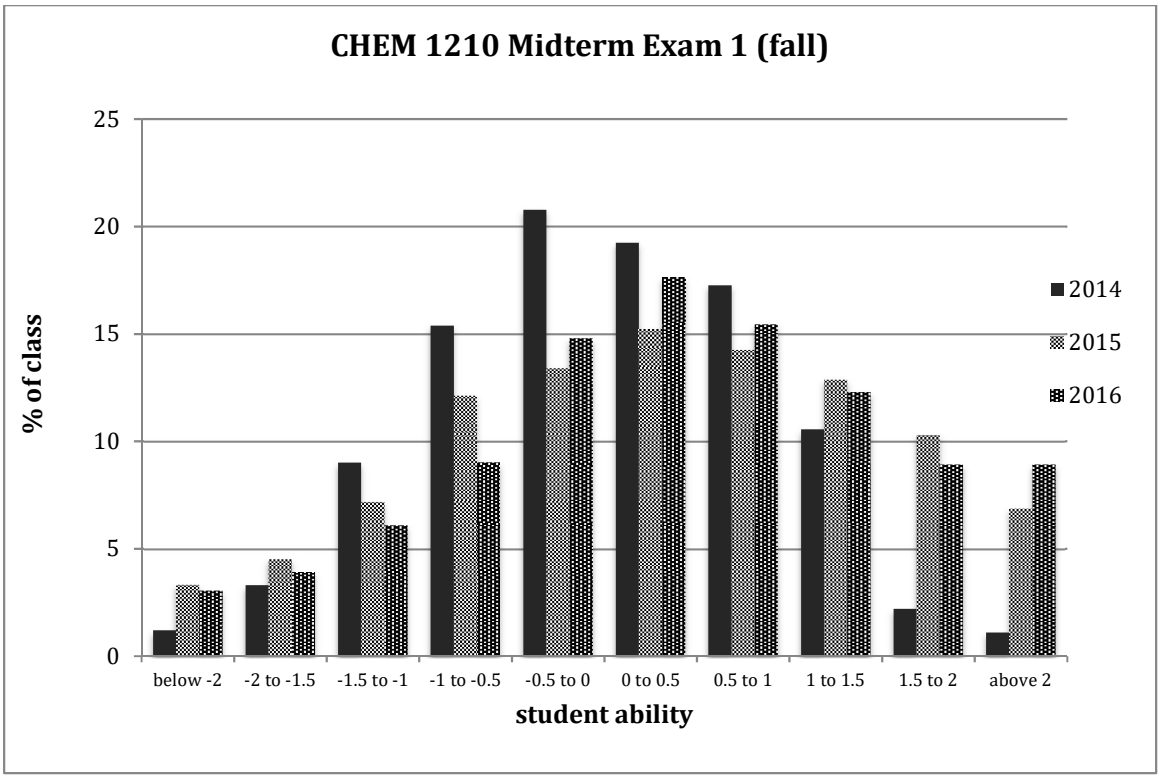
Table 4: Feedback that each student in Table 3 received based on their STA's. STA's above 1.5, between 1.5 and 0.5, between 0.5 and -0.5, -0.5 and -1.5 and below -1.5 were classified as well above average, above average, average, below average and well below average, respectively.

Student	**overall**	bonding	naming	quantum
I	well above average	well above average	well above average	well above average
II	above average	well above average	above average	well above average
III	above average	above average	average	well above average
IV	above average	above average	well above average	average
V	average	below average	above average	above average
VI	average	average	above average	well below average
VII	average	well above average	well below average	below average
VIII	below average	below average	below average	well below average
IX	below average	well below average	well below average	below average
X	below average	well below average	below average	well below average
XI	well below average	below average	well below average	well below average
XII	well below average	well below average	well below average	well below average

were changed or completely rewritten in order to prevent cheating and reflect the desires of the instructing professors. The exam scores were therefore equated using the questions in common as anchor points so that exam scores could be compared from one year to the next. Figures 5-9 show histograms of equated student abilities on each midterm exam as well as a tabulation of the means, standard deviations and number of students. The abilities from 2015 – 2016 and 2016 – 2017 are equated to the respective exam from 2014 – 2015. This effectively tells us how well a student from a subsequent year would have done if they'd taken the exam from 2014 – 2015.

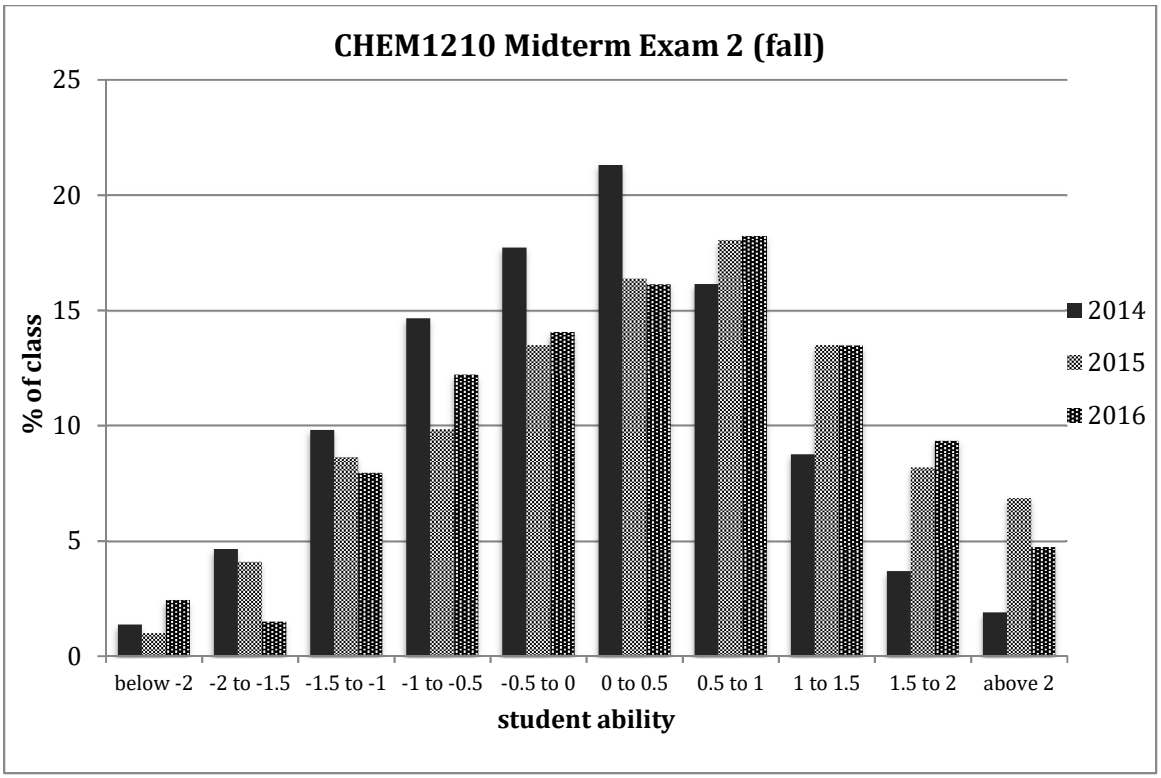
Improvement

As we hypothesized, we saw an improvement in student abilities after introducing the pretest system. The average student ability increased across all midterm exams in 2015 – 2016 compared to 2014 – 2015. While this difference was smaller than the standard deviations of those abilities, it was still statistically significant, with a p value less than 0.001 on all exams, due to large sample sizes. Student abilities remained largely unchanged from 2015 – 2016 to 2016 – 2017, confirming that the increase was not a one-year aberration. The difference in abilities between 2016 – 2017 and 2014 – 2015 was once again statistically significant with a p value less than 0.001 on all exams.
These averages suggest that student abilities increased overall as a result of our pretest system. Looking at the student ability histograms, we can examine more closely who among the students were benefitting most. On the first exam, Figure 5, the most noticeable difference between 2014 and 2015 is the shift of students from average ability to high ability. The proportion of students with an ability above 1.5, which we would



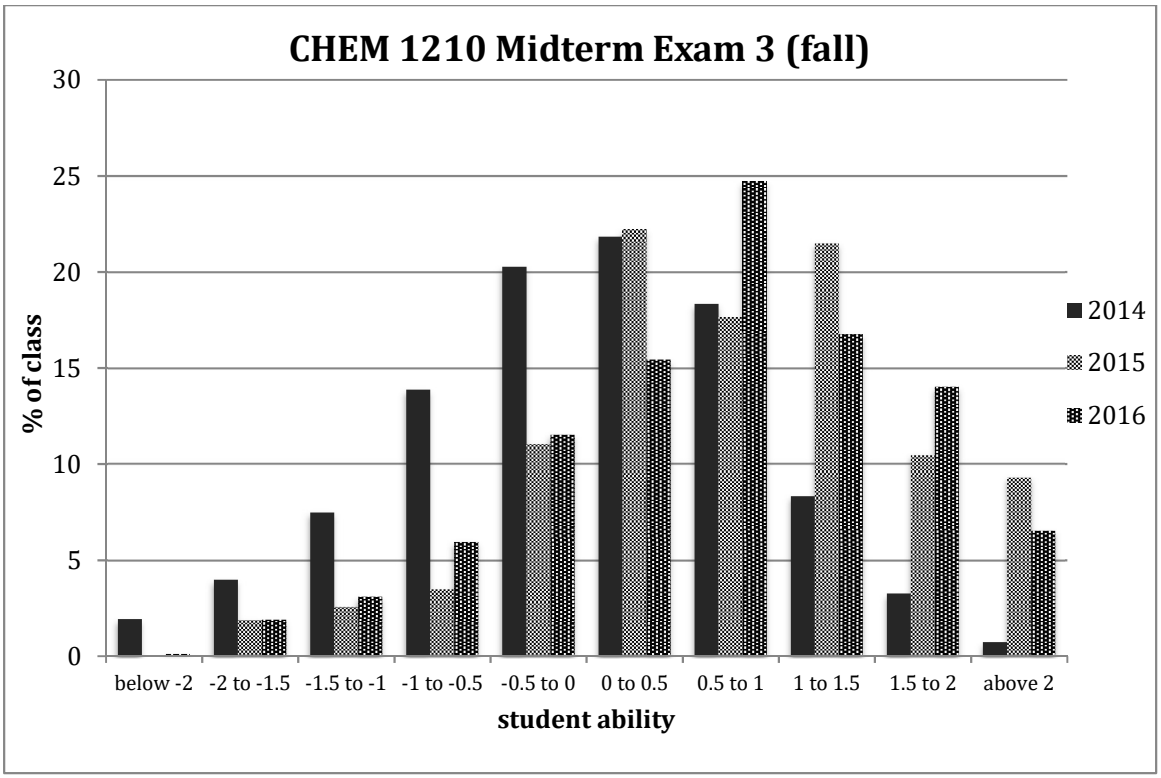
	2014 original	2015 equated	2016 equated
Average ability	0.00	0.27	0.36
Standard deviation	0.88	1.19	1.18
N	910	930	917

Figure 5: Comparison of CHEM 1210 Exam 1 results from 2014 to 2016. a) Histogram of student IRT abilities. The abilities from 2015 and 2016 are equated to the 2014 exam. b) A summary of the mean, standard deviation and number of student abilities in the above histogram.



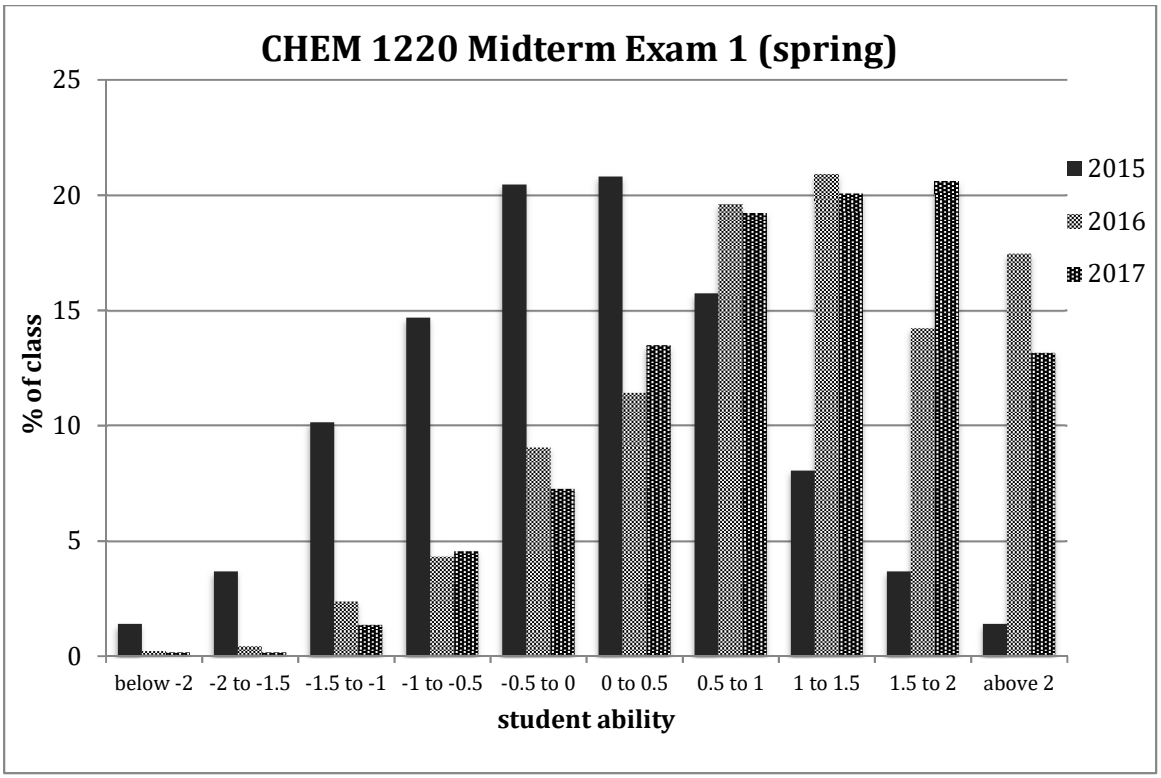
	2014 original	2015 equated	2016 equated
Average ability	0.00	0.33	0.30
Standard deviation	0.94	1.09	1.08
N	945	901	865

Figure 6: Comparison of CHEM 1210 Exam 2 results from 2014 to 2016. a) Histogram of student IRT abilities. The abilities from 2015 and 2016 are equated to the 2014 exam. b) A summary of the mean, standard deviation and number of student abilities in the above histogram.



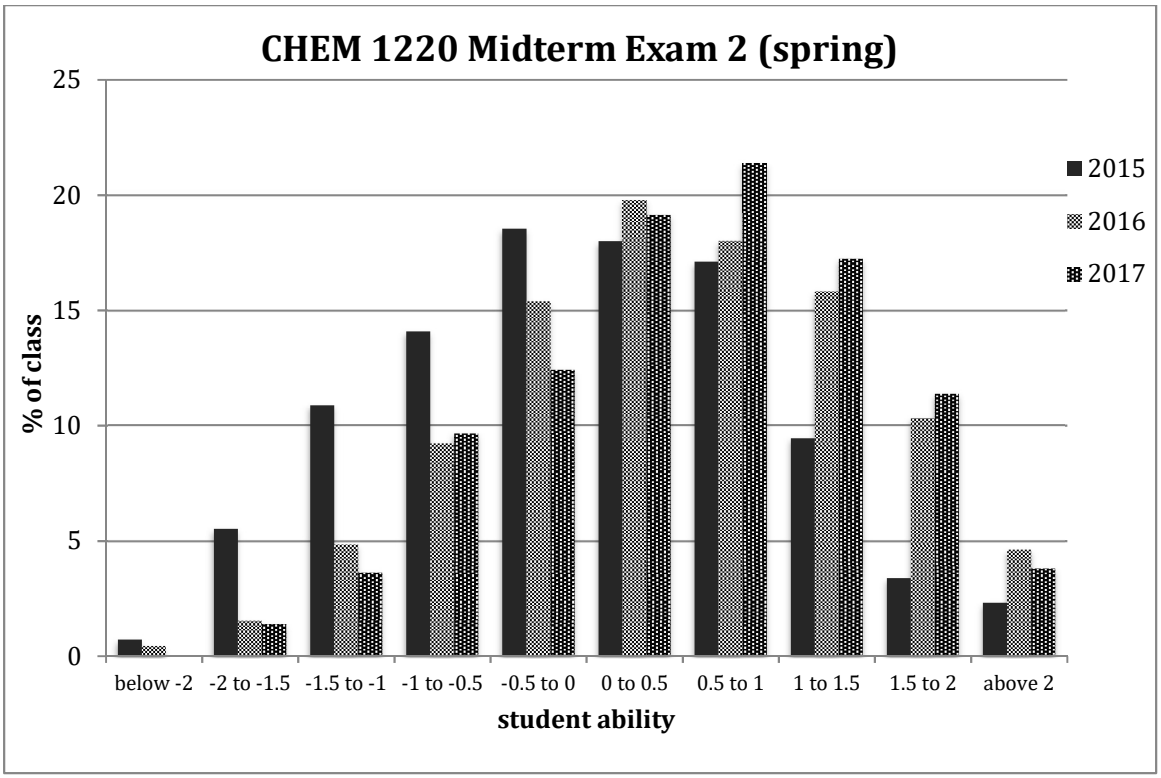
	2014 original	2015 equated	2016 equated
Average ability	0.00	0.75	0.68
Standard deviation	0.89	0.93	0.93
N	826	857	839

Figure 7: Comparison of CHEM 1210 Exam 3 results from 2014 to 2016. a) Histogram of student IRT abilities. The abilities from 2015 and 2016 are equated to the 2014 exam. b) A summary of the mean, standard deviation and number of student abilities in the above histogram.



	Spring 2015	Spring 2016	Spring 2017
Average ability	0.00	1.00	1.01
Standard deviation	0.92	0.98	0.88
N	572	461	590

Figure 8: Comparison of CHEM 1220 Exam 1 results from 2015 to 2017. a) Histogram of student IRT abilities. The abilities from 2016 and 2017 are equated to the 2015 exam. b) A summary of the mean, standard deviation and number of student abilities in the above histogram.



	Spring 2015	Spring 2016	Spring 2017
Average ability	0.00	0.45	0.54
Standard deviation	0.95	0.94	0.90
N	561	452	577

Figure 9: Comparison of CHEM 1220 Exam 2 results from 2015 to 2017. a) Histogram of student IRT abilities. The abilities from 2016 and 2017 are equated to the 2015 exam. b) A summary of the mean, standard deviation and number of student abilities in the above histogram.

consider well above average, increased from roughly three percent in 2014 to seventeen percent in 2015. At the same time though, the fraction of students scoring below -1.5 also increased, albeit less dramatically, from approximately five to eight percent. Perhaps this initial trial of the pretest system allowed some of the average students to identify and correct a single area of weakness, while the poorer students were overwhelmed by the negative feedback.

On the second midterm exam, the ability gains were more evenly distributed. The well below average students with abilities less than -1.5 decreased in number from six to five percent and those between -1.5 and -0.5 dropped from twenty four to eighteen percent of the class. By the time of the third midterm in 1210, almost sixty percent of the class had an above average or well above average ability and the well below average category had dropped to a mere two percent. This trend of increased abilities across the three exams of CHEM 1210, particularly among the lowest performing students, may propose that increased familiarity with the pretest system led to better outcomes. In other words, the students had to learn how to best utilize this study tool.

On the first exam of CHEM 1220, we see an even greater increase in student abilities from spring 2015 to 2016. In fact, the average student in 2016 had an ability of 1.0, putting them a full standard deviation above the respective students from 2015. At this point, only sixteen percent of students had an ability lower than 0. It is not surprising that scores continued to increase in the second semester; the vast majority of students enrolled in 1220 in any given spring semester had just taken 1210 the previous fall, so the students in spring 2016 were already well versed with the pretest system.

Interestingly, this pattern of ability improvements increasing in magnitude over

the course of the year came to a halt with the second and final midterm exam of 1220. Abilities were still significantly higher in spring 2016 than in 2015 but the average ability only increased by 0.45 compared to the 1.0 of the first exam. It is possible that this particular exam is less well suited to incremental improvements in individual areas since these topics are more interdependent on one another than most exam topics.

Effect of Feedback

During the first semester that the test review system was implemented, one of the three classes served as a control group. The students in this class still took the pretests but did not receive feedback informing them of their Likert abilities. We expected the students who received the feedback to exhibit higher abilities on the midterm exams relative to these students in the control class.

As shown in Table 5, however, we found that the two groups had roughly identical abilities on all three exams. While this does suggest that the feedback itself is not the be all and end all for improving student ability, several factors may have influenced this result. First, since the control group involved a single professor's class, the teaching ability of that professor relative to his colleagues would obviously make a difference. In addition, this class was not a random sampling of the overall population of students. The classes were each taught at different times during the day so the students who chose this particular time may have had inherently different abilities than the students in the other classes. Ideally, students would have been selected for the control group on a completely random basis; this was not possible in practice, however, due to limitations within the *Madra Learning* system and concerns about intra-class fairness.

Table 5: A summary of the mean, standard deviation and number of student abilities of the feedback (experimental) and control groups on each of the midterm exams of CHEM 1210 during Fall 2015. None of the differences is statistically significant.

	Exam 1			**Exam 2**			**Exam 3**		
	mean	Standard Deviation	N	mean	Standard Deviation	N	mean	Standard Deviation	N
Feedback	-0.043	0.916	591	0.024	0.934	724	0.008	0.910	692
Control	-0.008	0.946	179	-0.077	0.963	176	-0.028	0.928	165

In subsequent semesters, all students received feedback as part of the pretest system. Although we had not definitively shown that the feedback itself resulted in increased student abilities, it was anecdotally popular among the students, did not appear to decrease student abilities, and provided an important component in a further metacognitive study that we were undertaking.

If the feedback cannot solely explain the increase in ability due to the test review system, than what does? It seems likely that a combination of factors accounted for this improvement. Among these, the simple act of giving students an opportunity and incentive to practice looms large. More specifically, because we set up the pretests on *Madra Learning* to match the exams, the students were able to familiarize themselves with the testing system and probably felt more comfortable taking the exams in this way than students had previously. The pretests also gave students a chance to assess their own ability prior to the exam and provided a reality check to those students who were not performing as well as they had hoped. The individualized feedback may have played a role as well, although we are unable to quantify it at this time.

It would be an interesting further study to try to isolate this effect of the feedback. It would likely require a truly random division into experimental and control groups to see a significant difference if any exists. Another option would be to increase the number of questions on the pretests and exams to more accurately measure student topic abilities.

CONCLUSION

In our general chemistry program, we developed and instituted a system of pretests for students to complete prior to each midterm exam. These pretests were designed to allow students to practice the type and format of questions that they would see on the exam but with lower stakes. Furthermore, this system analyzed the results of the pretests and determined, using IRT, individual students abilities on each topic of the pretest. This had not been done previously at the individual student level.

Based on this analysis, our system then provided students with feedback, detailing their areas of strength and weakness. We expected that the pretests would result in an increase in student performance by the combination of providing students an opportunity to practice, allowing students to assess their own understanding of the material and giving them particular areas to focus on to make the largest improvements.

We assessed overall improvement by comparing the equated student abilities from exams during 2014 – 2015, the year before we'd implemented our pretest system, to the years after, 2015 – 2016 and 2016 – 2017. Student abilities were equated to allow comparison between the different examinations that were given each year. We found that equated student abilities increased significantly in 2015 – 2016 across all exams. This effect was most pronounced after several exams, suggesting that students needed an adjustment period to acquaint them with the new system.

During the first semester of the pretest system's implementation, we also sought

to determine the effect of the feedback within the overall system. One class of the three being taught that semester served as a control, with its students taking the pretests but not receiving feedback. The students in this class had abilities virtually identical to the students who had received feedback. This indicated that if the feedback had any effect, it was likely not a major one. However, because a different professor taught each class, it is difficult to say whether we were indeed isolating the effect of the feedback.

In any case, we developed a novel method for the determination of student topic abilities and integrated it into a system that gave students the opportunity to practice and receive detailed feedback, which as a whole resulted in increased student abilities.

REFERENCES

Özmen, H. Some Student Misconceptions in Chemistry: A Literature Review of Chemical Bonding. *J. Sci. Educ. Technol.* **2004**, *13* , 147-159.

Madra Learning LC. Retrieved from Madra Learning: https://app.madralearning.com accessed 2017**.**

An, X.;Yung, Y.-F. Item Response Theory: What It Is and How You Can Use the IRT Procedure to Apply It. SAS Institute Inc. 2014.

Ayala, R. J. *The Theory and Practice of Item Response Theory.* The Guilford Press: New York, 2009.

Bell, P.; Volckmann, D. Knowledge Surveys in General Chemistry: Confidence, Overconfidence, and Performance. *J. Chem. Educ.* **2011,** *88*, 1469–1476.

Childs, P. E.; Sheehan, M. What's difficult about chemistry? An Irish perspective. *Chem. Educ. Res. Pract.* **2009**, *10* , 204-218.

Chiu, M.-H.; Chou, C.-C.; Liu, C.-J. Dynamic Processes of Conceptual Change: Analysis of Constructing Mental Models of Chemical Equilibrium. *J. Res. Sci. Teach.* **2002**, *39*, 688-712.

Dunlosky, J.; Metcalfe, J. *Metacognition.* Los Angeles: Sage. 2009**.**

Johnson, M. S. Marginal Maximum Likelihood Estimation of Item Response Models in R. *J. Stat. Software.* **2007**, *20*, 1-24.

Johnstone, A. H. Chemical education research in Glasgow in perspective. *Chem. Educ. Res. Pract*. **2006**, *7* , 49-63.

Kruger, J.; Dunning, D. Unskilled and Unaware of It: How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-Assessments. *J. Pers. Soc. Psychol.* **1999**, *77* , 1121-1134.

McClary, L.; Talanquer, V. College Chemistry Students' Mental Models of Acids and Acid Strength. *J. Res. Sci. Train.* **2011**, *48* , 396-413.

Pınarbasi, T.; Canpolat, N. Students' Understanding of Solution Chemistry Concepts. *J.*

Chem. Educ. **2003**, *80* , 1328-1332.

Sanger, M. J.; Phelps, A. J. What Are Students Thinking When They Pick Their Answer? *J. Chem. Educ.* **2007**, *84* , 870-874.

Schurmeier, K. D.; Atwood, C. H.; Shepler, C. G.; Lautenschlager, G. J. (2010). Using Item Response Theory to Assess Changes in Student Performance Based Upon Changes in Question Wording. *J. Chem. Educ.* **2010**, *87* , 1268-1272.

Schurmeier, K. D.; Shepler, C. G.; Lautenschlager, G. J.; Atwood, C. H. (2011). Using Item Response Theory To Identify and Address Difficult Topics in General Chemistry. *ACS Symposium Series.* **2011**, *1074* , 137-176.

Tanaka, J.S. Multifaceted conceptions of fit in structural equation models. In K.A. Bollen & J.S. Long (Eds.), *Testing Structural Equation Models.* Newbury Park, CA: Sage. 1993.